Redundant Documents and Search Effectiveness

Yaniv Bernstein Justin Zobel
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V
Melbourne, Australia 3001

ABSTRACT

The web contains a great many documents that are *content-equivalent*, that is, informationally redundant with respect to each other. The presence of such mutually redundant documents in search results can degrade the user search experience. Previous attempts to address this issue, most notably the TREC novelty track, were characterized by difficulties with accuracy and evaluation. In this paper we explore syntactic techniques — particularly document fingerprinting — for detecting content equivalence. Using these techniques on the TREC GOV1 and GOV2 corpora revealed a high degree of redundancy; a user study confirmed that our metrics were accurately identifying content-equivalence. We show, moreover, that content-equivalent documents have a significant effect on the search experience: we found that 16.6% of all relevant documents in runs submitted to the TREC 2004 terabyte track were redundant.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Selection Process*

General Terms

Performance, Measurement

Keywords

Novelty, duplicate detection, search effectiveness

1. INTRODUCTION

Search engines are designed to present answers in a way that allows users to satisfy their information needs with minimum effort. An issue that interferes with this objective is data redundancy. If a document is effectively identical to documents that have already been presented, then it is unnecessary for the user to see it; at the most they may be interested in its location or in the fact that it exists. However, most mainstream information retrieval systems — and most methods for measuring them (van Rijsbergen 1979) — are based on the assumption that the relevance of each document is independent of any consideration of other documents in the result list.

Several techniques for managing repetition of topics in answer lists are based on the assumption that different documents may contain alternative presentations of the same information. For example, in Scatter-Gather (Hearst & Pedersen 1996) documents are grouped based on automatic clustering. Other approaches to reducing redundancy have been explored in the TREC novelty track (Harman 2002, Soboroff & Harman 2003), where the aim is to extract novel and relevant sentences from an ordered list of relevant documents. However, it is unclear whether it would be feasible to add any of the methods evaluated in the track to a standard search engine (Allan et al. 2003). Furthermore, the novelty approach suffers from the problem that novelty as a concept is difficult to define, recognize and evaluate.

In this paper we examine a more elementary and robust approach to reducing redundancy in search results. We define pairs of documents that contain the same information as each other as *content-equivalent*. In many cases content-equivalent documents can be removed from a result list with little or no negative consequence. Given knowledge of content-equivalence relationships within a collection, answer lists can be postprocessed to present duplicates as a single entry, reducing the presentation of redundant information to the user.

We explore syntactic techniques for identifying content-equivalent pairs. Our approach is deliberately conservative; in this application an error-prone algorithm may inadvertently conceal essential information from the user. Nonetheless, we show that our approaches are able to identify a large number of content-equivalent pairs on web collections. We examine document fingerprinting (Manber 1994, Brin et al. 1995, Heintze 1996, Broder et al. 1997), a technique that can rapidly analyze a collection to identify pairs of documents that share significant blocks of text. Using human assessors and a substantial set of document pairs of different degrees of syntactic similarity, we experimentally determine thresholds of similarity above which most documents are content-equivalent.

We experimentally explore the effect that content-equivalence amongst documents in a collection has on the search experience. Our results on the TREC GOV1 and GOV2 document collections show that not only are there many content-equivalent documents but that this large-scale duplication leads to significant redundancy in result lists returned by systems for real queries. We also evaluate the consequences of content equivalence for search and for evalu-

CIKM'05, October 31–November 5, 2005, Bremen, Germany.

ation of information retrieval systems. We find, surprisingly, a high level of inconsistency in the relevance judgements for the 2004 terabyte track; our results show that a good fraction of the judgements are wrong. Moreover, our results show — on the assumption that redundant information is irrelevant — that content-equivalent documents are dramatically affecting search effectiveness results. We show that removing content-equivalent documents from result lists has the potential to significantly improve effectiveness.

2. REDUNDANCY IN SEARCH RESULTS

The process of information retrieval can be explained in terms of a hypothetical user with a specific *information need* (van Rijsbergen 1979). The user's objective is to satisfy the information need; the role of the information retrieval system is to retrieve and present information in such a way that the information need can be satisfied with minimum expenditure of effort by the user. The expression of the information need by the user, the presentation of results by the system, and the interaction between the user and the system can all take many forms. However, in the domain of text search, most systems use variations on a single model, in which the user presents a query and the system returns a ranked list of documents that the user is able to browse. The aim is to structure the list such that all relevant documents in the collection are at the beginning. The user can minimize their effort by browsing the list in order until the information need is met.

The extent to which an information retrieval system achieves these goals is quantified by values such as recall and precision. Recall is a measure of the extent to which all the available relevant information has been presented to the user, while precision is a measure of the extent to which the effort required to fulfill the information need has been minimized. These measures simplify the model by assuming independence of relevance. That is, in all cases relevance is assumed to be a property exclusively of the relationship between a given document and an information need. The position of the document within a result list and its relationship to other documents in the list is not considered.

However, as we show later, removal of this assumption does affect measured performance. In practice, it is often the case that, although a document is in isolation relevant to the information need, by the time it has been viewed by the user it adds nothing new. It follows that for typical querying such a document should not be considered relevant.

For some information needs, the existence or location of redundant documents may be of value to the user, but such cases are probably rare; in most cases, it is the content itself that the user seeks. In other words a document, in order to be relevant, must not just be topical to the information need but must in addition contain an element of *novelty*.

The TREC novelty track (Harman 2002, Soboroff & Harman 2003) has been a forum for evaluating the performance of systems that attempt to promote novelty by removing redundant information from the result list. The track focused on a sentence-level retrieval task in which each sentence was returned only if it was relevant and contained some novel information given the sentences that preceded it.

There are several problems associated with the task defined by the novelty track. The first of these is assessment. The fact that the novelty of a sentence depends on every sentence that precedes it means that the task has had to be evaluated on a small, predefined set of documents presented in a specific order. Furthermore, in order to have a larger pool of relevant sentences on which to base the assessments, all documents in the list were strongly relevant to the query. However, this practice may have introduced a bias in the results. Allan et al. (2003) suggest that this common practice in novelty research of using only relevant documents could mean that existing results do not predict performance in more realistic search environments.

The consequences of system error in the novelty task are another cause for concern. Semantic concepts such as novelty — or, indeed, relevance — are notoriously difficult to reliably detect with a computer. This is reflected in the low recall and precision scores prevalent in most areas of information retrieval; high accuracy is not currently achievable. The problem is that, while inaccuracy is somewhat acceptable for relevance, it is less so for novelty. Because the systems aim to highlight only novel information, all other information — that which is considered to be redundant — is effectively discarded from the search result. Thus, misclassification of novel information as redundant could lead to critical information being overlooked.

Furthermore, the novelty track has demonstrated that not even human judges were able to agree on the novelty status of many sentences. Given the inherent difficulty of the task, and the further difficulty we have using computers to capture semantic properties such as novelty, it is unlikely in the near future that systems will be able to discern novelty with a sufficient degree of accuracy to be useful in any but the most specialized applications.

Finally, there is more to meeting a user's information need than just novelty; the issue of authority is also a consideration. If a user were to discover an unlikely fact (for example, that cold fusion has been achieved) from a single source, then they would be liable to doubt that information. However, if they came across that same information from several independent sources then they would be inclined to lend more weight to the truth of the statement. Thus, elimination of semantic redundancy may not always be advantageous.

For all of the above reasons, the techniques explored in the TREC novelty track are unlikely to gain much traction amongst production information retrieval systems.

3. CONTENT EQUIVALENCE

We refer to a pair of documents as *content-equivalent* if they convey the same information as each other. A functional interpretation is that an information consumer, having viewed one document, would gain no new information by viewing the other.

In developing a robust technique for detecting content-equivalent pairs, we restrict ourselves to detecting relationships between documents with high syntactic similarity — pairs typically referred to as duplicates or near-duplicates. In contrast to the task of analysing the semantic content of documents, purely syntactic analysis is relatively easy, as it relies only on the superficial structure of the data. This is a conservative approach; we may miss some content-equivalent pairs, but the probability of false positives is much lower than when using semantic techniques.

A further advantage of this approach is that using syntactic similarity as criterion largely resolves the issue of authority. In many cases, repetition of information from multiple independent sources is an important step towards meeting an information need. By contrast, duplication is rarely valuable: further copies of the same document add little to the user's confidence in the information. For most scenarios, a result list need not contain both documents from a content-

equivalent pair; one is sufficient, with perhaps an indication that the same document appears elsewhere.

It is trivially apparent that two identical documents are content-equivalent. However, documents that are not bytewise identical can still display enough syntactic similarity to be identified as content-equivalent. A common scenario in the TREC .gov data is the presence of two identical documents, one in HTML and the other converted to plaintext from PDF. These documents are identical on a word-by-word level but vary significantly in the way they are stored. In other cases, documents vary in their formatting and hyphenation. Another kind of case is where a document is updated without affecting most of the text, such as alternative forms of the same press release or different versions of the same policy document.

Duplication or near-duplication of documents has been noted as a special-case or extreme-case example of document redundancy (Zhai et al. 2003, Zhang et al. 2002). We show that document duplication and near-duplication is neither special nor extreme; rather, it is widespread and has a significant impact on search results. We describe two syntactic approaches to robust identification of content-equivalent documents below.

4. DETECTING EQUIVALENCE

We first define *retrieval equivalence*, an easy-to-compute restricted form of content equivalence motivated by the operation of search engines. A set of documents is retrieval-equivalent if, once they are canonicalized — stripped of formatting and other information that is not considered during indexing — they are identical. This means that the documents will be indistinguishable to a search engine at retrieval time. Retrieval-equivalent document sets provide a very strong assurance of redundancy, as their textual content is virtually identical.

The algorithm for detecting retrieval equivalence follows from the definition: documents in the collection are canonicalized and then hashed using the MD5 algorithm (Rivest 1992). The list of hash values is sorted and all sets of documents sharing an identical MD5 hash are considered to be retrieval-equivalent; the asymptotic cost is thus $O(n \log n)$ for n documents but in practice the dominant cost is reading the documents.

We experimented with six levels of canonicalization, where each level adopts all the measures of previous levels: whitespace normalized; tags removed; punctuation removed; case folded; stopwords removed; and words stemmed. Note that the possibility of hash collisions is negligible: MD5 is a 128-bit hash, meaning that the space of possible values is 2^{128}; on a collection of a billion documents the likelihood of a single collision is about 10^{-25}.

More general detection of syntactic similarity for non-identical documents is less straightforward. There are several methods available for determining syntactic similarity between documents in a text collection, such as the I-Match algorithm of Chowdhury et al. (2002) and relative-frequency techniques such as those of Hoad & Zobel (2003) and Shivakumar & García-Molina (1995). We choose however to use document fingerprinting techniques (Manber 1994, Brin et al. 1995, Heintze 1996, Broder et al. 1997). We do this because the technology has a proven record of application to large document collections (Broder et al. 1997, Cho et al. 2000, Fetterly et al. 2003), and because it can easily be tuned to capture different levels of duplication.

Document fingerprinting is based on a process whereby fixed-size document *chunks* — for example, word sequences of length eight — are compared to each other. The level of similarity between a pair of documents is determined by the number of chunks they have in common. This is an appealing measure, as identical documents have all their chunks in common, and the measured similarity degrades gracefully as the documents diverge as through revisions, deletions and insertions. Content-equivalent document pairs are expected to score highly using such schemes. Furthermore, given an appropriate chunk length, false positives are unlikely.

Retaining the full set of chunks for each document would consume a quantity of resources several times the size of the source collection. Most document fingerprinting techniques reduce resource consumption by using a heuristic selection function to choose which chunks to retain. Furthermore, most systems also hash each chunk so that it takes less space. While hashing introduces the possibility of false positives, the space of chunks is sufficiently sparse that they are rare.

Once chunks have been selected, they are inserted into a standard inverted index (Witten et al. 1999). For our application — in which we are interested in syntactic similarity between all pairs of documents in the collection — we can step through each postings list in the index and update accumulators for pairs of documents that co-appear in that list. Shivakumar & García-Molina (1999) describe a refinement of this approach that reduces costs associated with the quadratic expansion of postings list. We use this latter approach in our work.

The most significant difference between fingerprinting techniques is in their chunk selection process. Most selection techniques use simple heuristics, such as selecting chunks based on their hash value or some other superficial characteristic of the chunk. There are other more sophisticated techniques, such as the winnowing algorithm of Schleimer et al. (2003). However, in all cases these selection schemes are lossy; potentially valuable data is discarded, rendering the algorithms less reliable and less able to distinguish between levels of syntactic resemblance.

The SPEX selection algorithm (Bernstein & Zobel 2004) provides lossless chunk selection. It proceeds from the observation that a chunk that occurs only once in a collection does not affect the calculation of similarity scores, as the calculations are based on chunk co-occurrence. SPEX uses an iterative hashing algorithm to select only chunks that occur more than once in the current collection. In most collections, a great many chunks occur once only, so this scheme provides significantly reduced index sizes without the loss of accuracy associated with the schemes reviewed above. For these reasons we have chosen to use SPEX as the chunk selection algorithm for our experiments.

We used the S_3 measure (Bernstein & Zobel 2004) to compute resemblance between a pair of documents u and v:

$$S_3(u,v) = \sum_{c \in u \wedge c \in v} (\text{mean}\, \bar{u}, \bar{v})^{-1} \qquad (1)$$

The S_3 score measures the number of chunks shared by u and v as a proportion of the mean number of chunks in u and v. This score takes on values between 0.0 and 1.0, making it natural to interpret the score as a similarity proportion or percentage.

As we intend to use the S_3 score as a predictor of whether a pair of documents is content-equivalent, we need to establish the correlation between this algorithmically-generated score and the subjective semantic notion of content equiv-

alence. Having determined the nature of this correlation, we can then choose a threshold S_3 value such that a pair of documents scoring higher than this value is — with a high degree of reliability — content-equivalent.

5. ASSESSING EQUIVALENCE

As the S_3 similarity scores of pairs of documents range smoothly between 0 and 1, we used human assessors to determine an appropriate threshold for content equivalence. For this experiment a large number of document pairs with S_3 scores evenly distributed between 0.4 and 1 were retrieved from the QRELS collection described below; based on prior inspection, to make the best use of our assessors we decided to classify all pairs with lower scores as not content-equivalent. Participants were presented with a sequence of these document pairs in a random order, and the S_3 score was not revealed. The participants assigned each of these pairs into one of the following categories:

Level 3. The documents have **completely equivalent** content; any differences between the documents are trivial and do not differentiate them with respect to any reasonable query.

Level 2. The two documents are **conditionally equivalent**; with respect to any query for which both documents may be returned by a reasonable search engine, the documents are equivalent. Any query for which the documents are not equivalent would only return one or other of the documents.

Level 1. The documents have **nearly equivalent** content with respect to any query for which both documents may be returned by a reasonable search engine; for those queries where the documents are differentiated, the differentiation is minor.

Level 0. The documents are **not equivalent**; differences between the documents are significant enough to differentiate them with respect to reasonable queries.

We avoid defining the term 'reasonable' too precisely; a reasonable query is a best-effort attempt at expressing an information need that one might plausibly have. A reasonable search engine makes sensible use of available information in the documents in order to make a best-effort attempt at answering the information need expressed in a given query.

A group of 4 people participated in a pilot study in which 420 document pairs were analyzed. The study revealed that there were many pairs for which content was equivalent in all but very particular circumstances. For example, two documents may have the same body but different navigational links. Although the navigational links would not differentiate the documents with respect to most queries, some queries may directly address these navigational links. In such cases the documents would not be equivalent.

It was this observation that suggested the concept of conditional equivalence, in which the space of queries can be partitioned into two categories: those for which the two documents are not differentiated, and those for which only one or other of the documents would be returned by a reasonable search engine. In other words, no queries exist for which both documents may reasonably be considered relevant to that query and yet at the same time not be considered equivalent. The significance of conditional equivalence is that, within a given result list, documents of this class can be considered content equivalent; the presence of both documents in the result list implies that they are equivalent for that query.

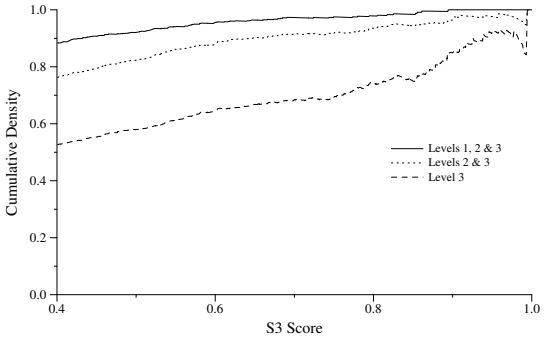

Figure 1: ***The cumulative proportion of document pairs judged to be at each equivalence level, for S_3 values from 0.4 to 1. Cumulation is from right to left.***

This change indicates how the definition of equivalence is far from absolute. We believe that the definitions that we have arrived at are useful for the particular application that we are investigating — web search — and that the relative consistency of responses we received from our assessors is an indication that the definitions correspond to real-world tasks. However, this does not mean that our definitions are necessarily suitable for a different set of assumptions or circumstances, and other definitions may be equally suitable for the web search domain.

For the main study, a group of 12 participants assessed a total of 964 document pairs for equivalence using the above criteria. Figure 1 shows the cumulative error rates as the S_3 threshold increases; that is, the values at a particular S_3 value indicate the level of accuracy for all documents at and above that score. The graph indicates a reasonably strong linear relationship between the S_3 score for a document and the likelihood that it is content-equivalent. Note the relatively low accuracy for level-3 document pairs. This result means that many document pairs identified by SPEX do have some point of difference (albeit a very small one in many cases). This suggests that it may not be prudent to purge documents from a collection at index-time, as there may be particular situations in which a given document may be needed despite its near-equivalence to other documents. However, the S_3 score shows high levels of accuracy for level 1 and level 2 document pairs.

The choice of a threshold is somewhat arbitrary given results such as these. We chose a threshold for which 95% of document pairs were classified as equivalent at level 1 or above. On the data collected, this threshold at which level 1 accuracy exceeds 95% is 0.58. The proportion of documents at this threshold that were categorized at level 2 or above was 87%. We use this threshold value to signify conditional content-equivalence.

Our results show that the S_3 measure is able to accurately identify content equivalence between documents. While not failure-proof, the method appears robust; given that the document in a content-equivalent pair that would be returned to the user is the document that is more highly ranked with respect to the query, and that the other document in the pair should be available to the user on request, the incidence of information loss is likely to be low.

Collection	Year crawled	Size (GB)	# documents
GOV1	2002	18.1	1,247,753
GOV2	2004	426.0	25,205,179

Table 1: ***Statistics for the GOV1 and GOV2 collections.***

	Clusters	Files	Duplicate Files
GOV1	130,343	397,713	267,370
GOV2	1,750,660	4,701,610	2,950,950

Table 2: ***Numbers of duplicate files removed from GOV1 and GOV2 during the original crawl.***

6. EQUIVALENCE IN THE .GOV DOMAIN

We make use of three document collections for these experiments: GOV1, GOV2, and QRELS, which are collections of web documents from the .gov domain created for TREC. The GOV1 and GOV2 collections are crawls of the .gov domain; see Table 1 for details. According to the TREC track information, some duplicate documents have already been removed from these collections. Table 2 shows the number of files removed. QRELS is described later.

For the GOV1 collection, we use the SPEX algorithm of Bernstein & Zobel (2004) to determine retrieval-equivalence and content-equivalence. For the GOV2 collection, we were only able to identify retrieval-equivalence. Although SPEX appears to be reasonably scalable, processing several gigabytes an hour, the 426 GB GOV2 collection is too large for SPEX on our hardware.

Equivalence in GOV1. We tested the six levels of retrieval equivalence on the GOV1 collection. We found that when only whitespace and HTML tags were removed, there were 21,840 sets of retrieval-equivalent documents for a total of 97,048 documents. When all further transformations were applied, this increased to 22,870 sets for a total of 99,227 documents. In other words, most of these documents became identical by virtue of having their HTML tags removed. Few additional documents were identified as a result of further transformations. In light of this, all further experiments applied all transformations, as a stricter definition of retrieval equivalence would not change the numbers much. In general, documents in these sets can be considered completely content-equivalent (level 3) and in many cases all but one of these documents can be completely eliminated from the collection prior to indexing.

When we use SPEX on the collection, we find a total of 215,314 documents that are participating in a content equivalence relationship, or 116,087 additional documents compared to retrieval-equivalence only. In total this means that 17.3% of documents were non-unique within this relatively small collection, already a relatively high figure. As collections grow larger and more comprehensive, one would expect this proportion to grow.

We remarked above that the error rate for level 3 equivalence was too high to recommend pruning the collection at index time. Nonetheless, keeping a record of content-equivalence relationships within the collection enables efficient modification of result lists at query time, and gives an indication of the overall level of redundancy in the collection.

Retrieval equivalence in GOV2. Our analysis of retrieval equivalence in the GOV2 corpus found 865,362 retrieval-equivalent sets consisting of a total of 6,943,000 documents. Thus, a total of 6,077,638 documents in the collection are entirely redundant for retrieval purposes, in addition to the 2,950,950 duplicate documents that had already been removed at crawl time. These 6,077,638 documents represent nearly 25% of the documents in the official corpus, which is consistent with previous investigations of text duplication in web crawls (Broder et al. 1997, Fetterly et al. 2003), which reported figures in the range of 25%–30%. We speculate, based on our results for GOV1, that further large numbers of documents will be redundant according to the content-equivalence measure.

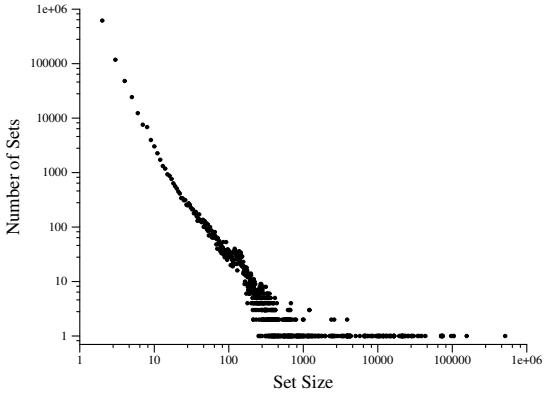

Figure 2: ***Distribution of retrieval-equivalent set sizes in GOV2, on log-log axes.***

The results on GOV2 show several extremely large sets of documents that are retrieval-equivalent. The largest set encompasses 512,030 documents — 2% of all the documents in the GOV2 collection. Figure 2 shows the frequency of occurrence of sets of various sizes. The linear character of the distribution on double-log axes suggests a power-law distribution. This is consistent with the results of Fetterly et al. (2003), with the graph showing a high degree of similarity to graphs constructed from a crawl of 150,000,000 documents from the general web. There is even a similar curious artefact at set sizes of about 100, which Fetterly et al. attribute to mirroring. There were many large sets that contained reasonably information-rich documents. An example of a document that occurred 300 times is in Figure 3.

7. EQUIVALENCE IN SEARCH RESULTS

The TREC terabyte track (Clarke et al. 2004) ran for the first time in 2004. The terabyte track is intended as a forum for assessing the retrieval performance of search engines on a far larger (and more consistent) web collection than had previously been available to the general research community.

The data set used for evaluation for the 2004 terabyte track consists of two components: the collection itself — 426 GB of data — and a set of 50 topics that define queries on the collection. Participants were required to run the set of queries on the collection using their search system and, for each query, submit the top 10,000 results as ranked by their software. In order to assess precision and recall, the top 100 documents for each query from each run were added to an assessment pool. The documents in the assessment pool were given one of three relevance values by an expert human judge, corresponding to not relevant, relevant and highly relevant. For the purposes of official assessment, the

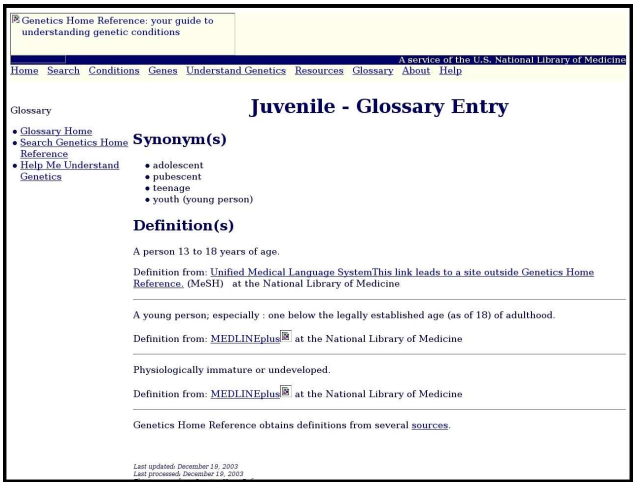

Figure 3: ***A document that appears 300 times in the GOV2 collection.***

latter two were treated as equivalent. Documents that were not in the pool were assumed to be non-relevant.

We wished to examine the occurrence patterns of content equivalence amongst documents that appeared in the relevance judgements for the 2004 terabyte track. To this end, we created a subcollection of the GOV2 collection, the QRELS collection, consisting exclusively of the human-assessed documents for the 50 queries of the GOV2 corpus. It consists of 58,078 documents and is 2.8 GB, or nearly one-sixth of the total size of GOV1.

There are several reasons why this subset of documents is worth investigating. Document access patterns are extremely skewed, such that some documents are accessed extremely frequently and others rarely or not at all (Garcia et al. 2004). For example, the documents that appeared in the largest sets discussed in Section 6 would be relevant to few queries and as such accessed rarely. The documents in the QRELS collection were retrieved in response to the carefully formulated TREC topics and as such have demonstrated their utility. The signal-to-noise ratio is expected to be much higher in the QRELS than amongst the collection as a whole. We note that the average size of a document in QRELS is 49.3 KB, or 2.8 times larger than the collection average of 17.7 KB. This is indicative of the higher information content of these documents.

Furthermore, the QRELS documents have been assessed for relevance by expert human judges. By examining the assessed documents for each query, we can analyze the prevalence of content equivalence amongst documents that are judged relevant. Finally, the QRELS documents allow us to observe the incidence of content-equivalent documents that have been inconsistently classified by the judges. This gives some indication of the overall quality of relevance judgements for the queries in this collection.

A total of 23.9% of all documents in the collection were found to exhibit content equivalence with at least one other document, but for some queries the figure was far higher. Only a minority of the content-equivalent documents are retrieval-equivalent; this shows that content equivalence has a real effect on search results.

Figure 4 shows the degree of document redundancy — the proportion of documents that can be eliminated for a given query because they can be represented by an equivalent document — on a per-topic basis. The lower histogram shows the percentage of all relevant documents for each topic that are redundant, while the upper histogram displays the same data for irrelevant documents. (Topic 703 was not included in the final TREC judgements, accounting for the absence of results for that query.) Taken across all queries, the mean percentage of redundant documents amongst the relevant set was 16.6%, while the mean amongst the irrelevant set was 14.5%. If we consider only retrieval-equivalent documents, mean redundancy amongst the relevant and irrelevant sets is 2.3% and 4.0% respectively.

We note that the observed incidence of content equivalence in the QRELS collection is lower than in the GOV2 collection as a whole. This is not surprising, as much of the content equivalence in the GOV2 collection occurred on low-value documents, such as the server-generated error pages and search forms that occurred in the large sets discussed above. Despite this, redundant documents still represent a significant proportion of all documents in both the relevant and nonrelevant judged sets of the QRELS collection. The fact that, on average, 16.6% of relevant documents are redundant means that the user will in many cases be viewing relevant documents that they have seen before. These documents, though relevant to the topic, are no longer relevant to the user's information need as they are entirely lacking in novelty. This suggests that effectiveness figures calculated for the terabyte track runs could be significant overestimates of the user experience, and that the user experience could be substantially improved by removing instances of redundant documents from result lists.

Inconsistency in relevance judgements. Many of the pairs flagged as content-equivalent contained inconsistent relevance judgements; one document was judged relevant while the other was judged irrelevant. Equivalent documents should not vary with respect to their relevance; thus, either the pairs were incorrectly classified as content equivalent or one of the documents in the pair was erroneously judged. This provides a convenient opportunity to assess the overall consistency of relevance assessment for this collection.

All documents that were connected by content-equivalence relationships were aggregated into a single group. Any group in which not all documents received the same relevance judgements were identified as containing an inconsistency. We manually examined 20 randomly selected groups that had been inconsistently judged and found that in all cases the judgements ought not have disagreed.

In total across all queries there were 465 groups where judgements were inconsistent. Within these groups, 791 documents were judged relevant and 681 were judged non-relevant. Assuming — for the sake of argument — that the majority of judgements in each groups was correct, on average across all groups 3.8% of the judgements (522 of 13854) were incorrect. If we consider only the set of *potentially relevant* documents — that is, the set of all documents that are either relevant or form part of an inconsistent group — we have 522 erroneous judgements out of 4,013 documents, an error rate of 13.0%.

This evidence seems incontrovertible: the same judge has inconsistently assessed identical or near-identical documents in many cases. We have no evidence of assessment quality for the documents that do not have a content equivalence relationship with any other documents, but there is no reason to suppose that accuracy for these other documents is any higher.

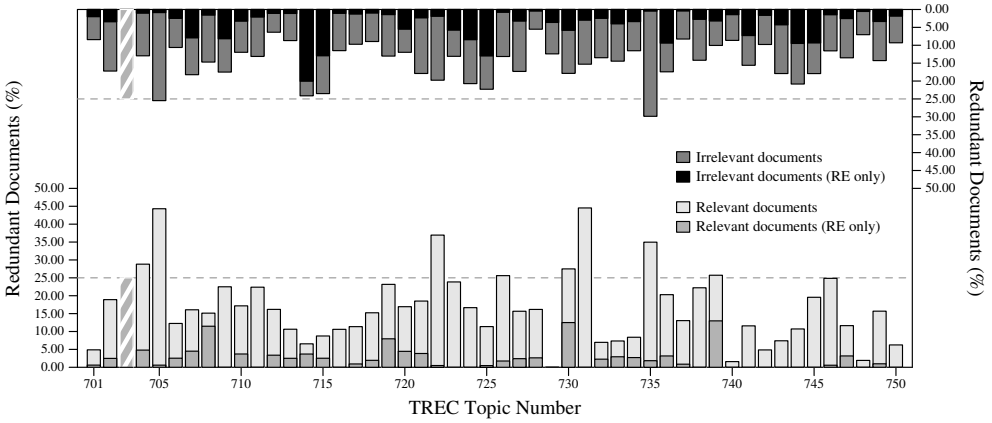

Figure 4: ***The proportion of redundant documents for each topic amongst the relevant (lower graph) and non-relevant (upper graph) judged pools. Note that query 703 is absent from the TREC judgements.***

Equivalence and search performance. Document redundancy has a negative effect on the effectiveness of search. In this section we quantify the extent of this effect on the 70 official runs submitted to the 2004 TREC terabyte track.

We reiterate the novelty principle, discussed earlier: a document — though relevant in isolation — should not be regarded as relevant if it is the same as a document the user has already seen. In order to model this principle, we modified the official judgements for each run so that documents appearing in a result list after another document with which they were content-equivalent were marked as irrelevant, regardless of their judged relevance status.

In order to discount the effect of poorly performing runs we adopted the methodology used by Voorhees & Buckley (2002) and Sanderson & Zobel (2005) and discarded the bottom 25% of runs, leaving us in this case with 54 runs. We used the `trec_eval` tool for evaluation and recorded the mean average precision (MAP) (Buckley & Voorhees 2000) for each of the runs.

Evaluating the TREC runs as submitted resulted in an average MAP across the runs of 0.201. When the novelty principle was modelled as described above, the average MAP fell to 0.161, a relative reduction in MAP of 20.2%. This is a substantial difference, and demonstrates that the assumption of independent relevance is inflating effectiveness scores, and that redundant documents almost certainly have a significant impact on the user experience of search. In a scenario when we are exploring the effect of eliminating redundant documents, it is clear that 0.161, not 0.201, is the correct baseline for comparison.

To simulate an information retrieval system that is aware of content-equivalent documents, we modified the runs so that documents appearing after another document with which they were content-equivalent were removed from the result list. The average MAP increase to 0.186, a relative 16.0% improvement in MAP compared to our baseline, demonstrating that an equivalence-aware retrieval system is able to substantially improve the user's search experience.

That the improvement is, by construction, observed in every query with redundant answers and that the degree of improvement is closely coupled to the degree of redundancy observed in the QRELS collection does nothing to invalidate this result. Again, we note that redundant answers can add no weight to a user's confidence in information. We also note that this improvement is obtained on the same underlying collection and on the same set of redundancy-modelled relevance assessments.

These results are presented in graphical form for each of the individual runs in Figure 5. In all cases the official MAP result overstates the effectiveness of the run. Interestingly, the improvement from removing duplicates was greater for runs that showed overall better effectiveness. This is an interesting phenomenon that will be studied in future work.

8. CONCLUSIONS

We have explored a class of similarity we call content equivalence; documents are content-equivalent if they contain the same information as each other. We contend that the presence of content-equivalent documents in a result list is not in general of benefit to the user. Document fingerprinting was shown via a user study to be a robust method for identifying content-equivalence between documents. Using our method, we found that over 17% of documents in GOV1 were non-unique under content equivalence, more than double the figure of the stricter retrieval equivalence. Almost 25% of GOV2 was redundant under retrieval equivalence; our analysis of judged documents and extrapolation from GOV1 suggests that as many documents again may be non-unique under content equivalence, suggesting a high degree of document redundancy in this larger collection.

We also showed that content-equivalence has a significant impact on actual search results from a wide variety of search methodologies. Purging content-equivalent documents from results lists improves novelty-based MAP on the TREC terabyte queries by an average of 16.0% over results generated by 54 different ranking algorithms. Disturbingly, our study exposes a significant degree of inconsistency in the human relevance judgements used for evaluating the performance of search algorithms in the TREC terabyte track.

Acknowledgements

This work was supported by the Australian Research Council and by an RMIT VRII grant.

9. REFERENCES

Allan, J., Wade, C. & Bolivar, A. (2003), Retrieval and novelty detection at the sentence level, *in* 'Proc. ACM SIGIR conference', ACM Press, pp. 314–321.

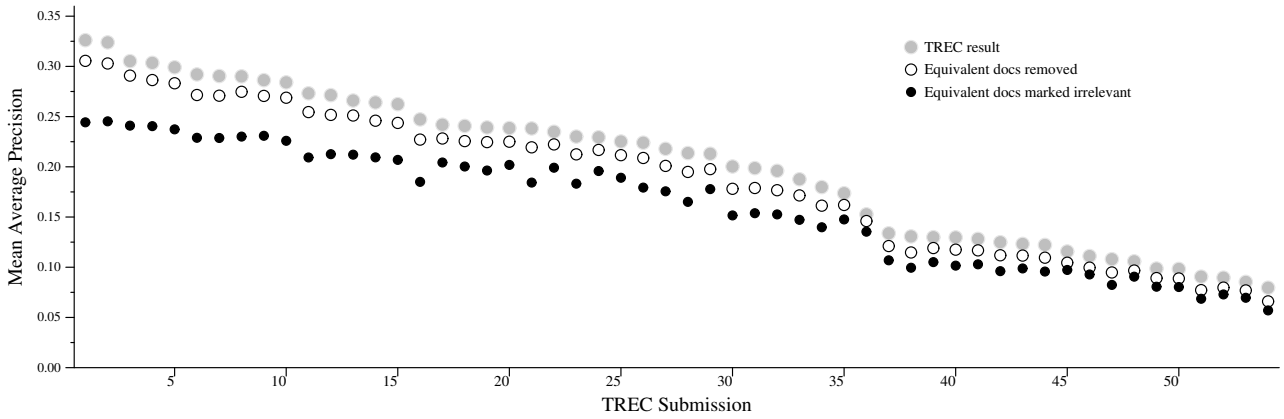

Figure 5: ***MAP on TREC queries 701-750 for the 54 most successful runs submitted to the 2004 terabyte track. The first column is the official MAP result, the second column shows MAP if subsequent content-equivalent documents are marked nonrelevant, and the third column shows MAP with the same documents removed from the run.***

Bernstein, Y. & Zobel, J. (2004), A scalable system for identifying co-derivative documents, *in* 'Proc. String Processing and Information Retrieval Symposium (SPIRE)', Springer, pp. 55–67.

Brin, S., Davis, J. & García-Molina, H. (1995), Copy detection mechanisms for digital documents, *in* 'Proceedings of the ACM SIGMOD Annual Conference', pp. 398–409.

Broder, A. Z., Glassman, S. C., Manasse, M. S. & Zweig, G. (1997), 'Syntactic clustering of the Web', *Computer Networks and ISDN Systems* **29**(8-13), 1157–1166.

Buckley, C. & Voorhees, E. M. (2000), Evaluating evaluation measure stability, *in* 'Proc. ACM SIGIR conference', ACM Press, pp. 33–40.

Cho, J., Shivakumar, N. & Garcia-Molina, H. (2000), Finding Replicated Web Collections, *in* 'Proc. ACM SIGMOD Conference', pp. 355–366.

Chowdhury, A., Frieder, O., Grossman, D. & McCabe, M. C. (2002), 'Collection statistics for fast duplicate document detection', *ACM Transactions on Information Systems (TOIS)* **20**(2), 171–191.

Clarke, C., Craswell, N. & Soboroff, I. (2004), Overview of the TREC 2004 Terabyte Track, *in* 'Proceedings of the 13th Text REtrieval Conference (TREC 2004)'.

Fetterly, D., Manasse, M. & Najork, M. (2003), On the Evolution of Clusters of Near-Duplicate Web Pages, *in* 'Proceedings of the 1st Latin American Web Congress', IEEE, pp. 37–45.

Garcia, S., Williams, H. E. & Cannane, A. (2004), Access-ordered indexes, *in* 'Proc. 27th conference on Australasian computer science', pp. 7–14.

Harman, D. (2002), Overview of the TREC 2002 Novelty Track, *in* 'The Eleventh Text REtrieval Conference (TREC 2002)'.

Hearst, M. A. & Pedersen, J. O. (1996), Reexamining the cluster hypothesis: scatter/gather on retrieval results, *in* 'Proc. ACM SIGIR conference', ACM Press, pp. 76–84.

Heintze, N. (1996), Scalable Document Fingerprinting, *in* '1996 USENIX Workshop on Electronic Commerce'.

Hoad, T. C. & Zobel, J. (2003), 'Methods for Identifying Versioned and Plagiarised Documents', *Journal of the American Society for Information Science and Technology* **54**(3), 203–215.

Manber, U. (1994), Finding Similar Files in a Large File System, *in* 'Proceedings of the USENIX Winter 1994 Technical Conference', pp. 1–10.

Rivest, R. (1992), 'The MD5 Message-Digest Algorithm'. RFC 1321.

Sanderson, M. & Zobel, J. (2005), Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability, *in* 'Proc. ACM SIGIR conference', pp. 162–169.

Schleimer, S., Wilkerson, D. S. & Aiken, A. (2003), Winnowing: local algorithms for document fingerprinting, *in* 'Proc. ACM SIGMOD conference', ACM Press, pp. 76–85.

Shivakumar, N. & García-Molina, H. (1995), SCAM: A Copy Detection Mechanism for Digital Documents, *in* 'Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries'.

Shivakumar, N. & García-Molina, H. (1999), Finding Near-Replicas of Documents on the Web, *in* 'WEBDB: International Workshop on the World Wide Web and Databases, WebDB', Springer-Verlag.

Soboroff, I. & Harman, D. (2003), Overview of the TREC 2003 Novelty Track, *in* 'The Twelfth Text REtrieval Conference (TREC 2003)', pp. 38–53.

van Rijsbergen, C. J. (1979), *Information Retrieval*, Butterworth-Heinemann.

Voorhees, E. M. & Buckley, C. (2002), The effect of topic set size on retrieval experiment error, *in* 'Proc. ACM SIGIR conference', ACM Press, pp. 316–323.

Witten, I. H., Moffat, A. & Bell, T. C. (1999), *Managing Gigabytes: Compressing and Indexing Documents and Images*, Morgan Kauffman.

Zhai, C. X., Cohen, W. W. & Lafferty, J. (2003), Beyond independent relevance: methods and evaluation metrics for subtopic retrieval, *in* 'Proc. ACM SIGIR conference', ACM Press, pp. 10–17.

Zhang, Y., Callan, J. & Minka, T. (2002), Novelty and redundancy detection in adaptive filtering, *in* 'Proc. ACM SIGIR conference', ACM Press, pp. 81–88.